Prévision spatio-temporelle de la criminalité à Chicago

Louis Pujol et Rémi Coulaud

19 février 2019

Table des matières

1 Introduction

Certains départements de police aux États-Unis utilisent des outils de prévision comme aide à la décision. Pourtant, nous ne sommes pas dans le film de science fiction "Minority Report" où la police est capable d'anticiper un crime et de le stopper avant qu'il soit commis. L'objectif des outils développés pour la prévision de crime est de donner à la puissance publique un moyen d'évaluer d'une part les lieux à risques et d'autre part de prédire le volume de crime commis à ces différents endroits. Précisons que la traduction française du mot crime devrait être infraction. En effet, en droit français le crime est l'infraction la plus grave parmi la contravention, le délit et le crime. Cependant dans le reste du rapport, nous nous autorisons cette traduction abusive du terme anglais. La modélisation de la criminalité a longtemps résisté aux statisticiens comme le note John.V.Pepper dans son article de 2007 [8]. Aujourd'hui plusieurs facteurs permettent d'espérer une amélioration de la qualité des prévisions. D'une part, l'accessibilité en continu des données permet d'utiliser des méthodes d'apprentissage par renforcement. D'autre part l'open source de certains jeux de données sur la criminalité permet à une communauté plus large de se saisir du problème comme le montre la popularité de ces jeux de données sur la plateforme Kaggle.

Il faut toutefois bien avoir conscience que la criminalité est un phénomène complexe faisant interagir des facteurs sociaux, économiques et politiques. Derrière le mot criminalité, nous regroupons un certains nombre de crimes divers et variés allant d'une infraction routière à un homicide. Nous touchons là un phénomène sensible comme le montre un des derniers numéro du journal signifiance sur l'utilisation des données dans les tribunaux [11]. Ce qui nous amène à nous demander pourquoi prédire le nombre de crimes. Ainsi, Wilpen Gorr *et al.* dans leur article de 2003 [5] montrent que la prédiction du nombre de crimes peut être faite à plusieurs échéances pour servir différents objectifs : à court terme (pour des redéploiements tactiques), à moyen terme (pour une ré-allocation des ressources) ou à long terme (pour orienter la politique sécuritaire de la ville).

D'un point de vue plus technique, nous verrons que la criminalité n'est pas seulement un problème temporel mais aussi un problème qui dépend de la spatialité dans la ville. Ainsi, nous cherchons, comme l'ont fait Wang *et al.* en 2016 [12] et Alex Reinhart et Joel Greenhouse en 2018 [9] à exploiter dans notre modélisation les caractéristiques spatio-temporelles de la criminalité. Il existerait une forme de persistance au cours du temps des crimes dans certains lieux ([2], [9]). Dans la perspective de prédire la criminalité dans une ville, il n'est pas seulement nécessaire d'obtenir de bonnes prédictions mais il est aussi nécessaire de comprendre les déterminants de la criminalité dans un lieu donné à l'aide d'une analyse multivariée. Enfin, il est avéré dans la littérature que la criminalité est un phénomène saisonnier comme le montre cette étude mainte fois cité de Gerhard J. Falk de 1952 [3] et l'article plus récent de Wilpen Gorr *et al.* [5]. Selon certains auteurs cela peut s'expliquer par le climat comme le montrent Simha F.Landau et Daniel Fridman en 1993 [7]. Nous allons maintenant présenter le jeu de données.

1.1 Présentation des données initiales

Dans cette première partie, nous présentons rapidement les données à notre disposition.

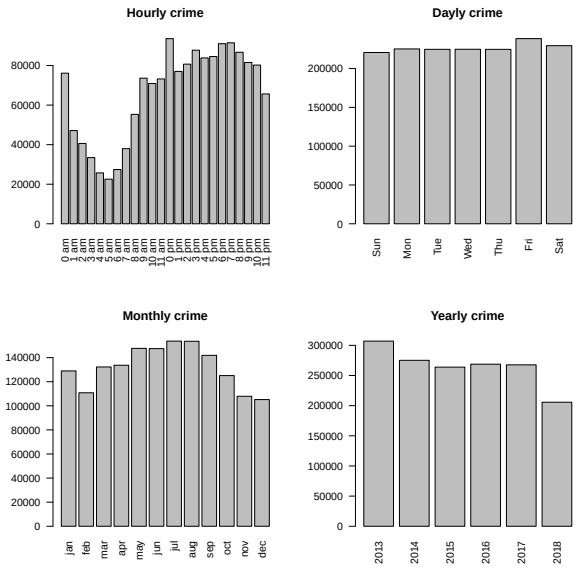

FIGURE 1 – Description temporelle des crimes

Le premier constat en analysant le nombre de crime agrégé par jour est la criminalité n'a pas une saisonnalité hebdomadaire très prononcée . Par contre nous constatons une saisonnalité annuelle forte. En effet, le pic haut a lieu durant l'été tandis que le creux de la criminalité est en hiver. Nous remarquons une tendance à la décroissance au cours des années du nombre d'infractions. Tendance beaucoup plus prononcée pour le trafic de drogue que pour les vols. Enfin, nous constatons comme Marcus Felson et Erika Poulsen en 2003 [4] que la criminalité a une saisonnalité journalière spécifique : le pic bas est à 5 heures du matin et le pic haut est vers 5 heures de l'après-midi. Nous confirmons ce résultat sur notre jeu de données grâce au graphique de la figure 1. Cette saisonnalité n'est pas la même si nous prenons en compte la spécificité des crimes : trafic de drogue ou vols. Ces remarques nous permettrons ensuite lors de la modélisation de chercher à capturer ces phénomènes de tendance et de saisonnalité à l'aide de variables temporelles.

Deuxièmement, nous avons à disposition des indicateurs spatiaux comme la localisation de chaque crime en coordonnées GPS ainsi que le nom de la rue et du quartier dans lequel il a eu lieu. Ces données nous permettent de représenter la répartition des crimes en 2017 sur la carte de Chicago (figure 2).

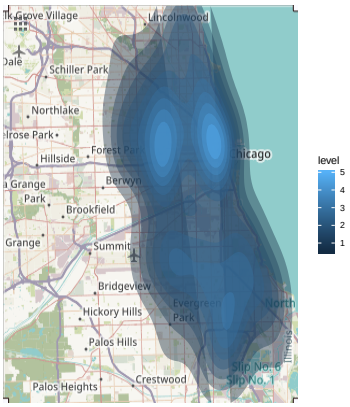

FIGURE 2 – Représentation spatiale de tous les crimes commis en 2017

Troisièmement, nous avons à disposition des variables concernant le crime lui même. On a d'abord une information concernant le type de crime. On peut distinguer quatre grandes catégories que sont les vols, les cas de violence sur personne, les dégradations matérielles et le trafic de drogue.

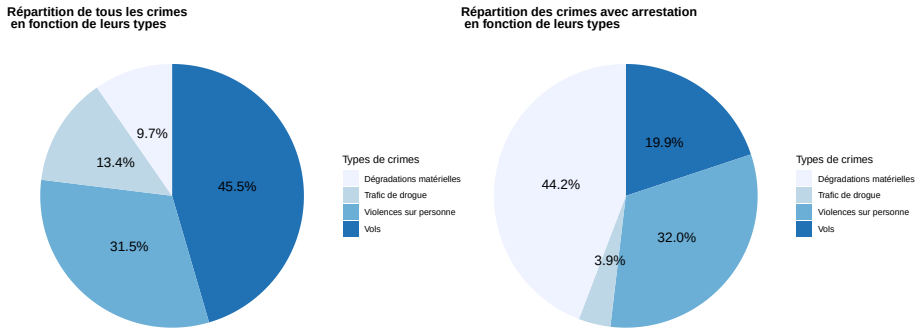

FIGURE 3 – Répartition des crimes selon le type

On sait également s'il s'agit d'un crime domestique et si le coupable a été arrêté. On remarque que le taux d'arrêt est très élevé pour le trafic de drogue, alors que beaucoup de vols sont signalés sans que le coupable soit arrêté. On comprend ce phénomène, une victime de vol ira naturellement porter plainte tandis que le trafic de drogue ne sera pas forcément dénoncé par des témoins, et les infractions répertoriées ne concernent que les trafiquants ou les consommateurs arrêtés. On remarque également sur la figure 4 que les violences sur personnes se produisent principalement dans l'entourage proche de la victime.

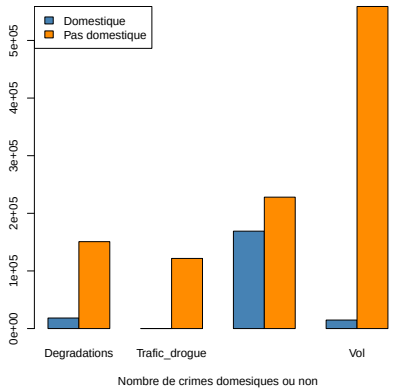

FIGURE 4 – Distribution des crimes domestiques par type de crime

1.2 Objectif et travail réalisé

L'objectif de ce projet est de prédire, sur une période d'un mois, le nombre de crimes ayant lieu dans la ville de Chicago chaque jour. Cependant, nous pensons qu'il est nécessaire d'aller plus loin que l'objectif de prévision et de chercher à comprendre et à expliquer les facteurs exogènes influant sur la criminalité. Nous avons pensé notre travail et le présent rapport comme pouvant être utile sous deux aspects à la ville de Chicago.

Premièrement, un outil performant de prévision du nombre de crimes commis chaque jour pourrait leur être utile dans la gestion des effectifs de police, si nous faisons l'hypothèse que le nombre de crime peut être réduit par une présence policière accrue. Deuxièmement, une mise en relation détaillée entre la criminalité et certaines composantes sociales et économiques pourrait permettre d'orienter des politiques publiques ciblées en direction de certains quartiers.

Pour répondre à l'objectif de prévision, nous allons construire des prédicteurs entraînés sur une certaine plage temporelle et évaluer leur performance de prédiction sur le mois suivant cette période. Afin de construire ces prédicteurs, nous serons amenés à introduire de nouvelles variables, issues d'autres jeux de données. Nous essayerons de comprendre comment ces variables sont liées au phénomène du crime à l'aide d'analyses descriptives et nous comparerons nos observations avec celles qu'ont pu formuler les auteurs ayant travaillé sur le sujet.

Nous commencerons par étudier qualitativement des variables spatiales, caractérisant la richesse et l'activité des différents quartiers de la ville de Chicago. Nous décrirons ensuite les modèles prédictifs envisagés ainsi que les variables temporelles introduites pour enrichir ces modèles. Puis, nous présenterons un algorithme d'agrégation spatiale de modèles que nous mettrons en œuvre sur nos données et nous analyserons qualitativement la sortie. Enfin nous discuterons de la mise en place d'une stratégie d'agrégation d'experts dans l'optique d'améliorer le score de prévision.

2 Analyse de la répartition spatiale des crimes à Chicago

Le jeu de données étudié comporte des informations sur la localisation et la nature des crimes commis dans la ville de Chicago. Dans cette partie nous tâcherons d'analyser et d'exploiter au mieux ces données. Après avoir mis en évidence l'irrégularité de la répartition des crimes sur le territoire, nous l'expliquerons à l'aide d'indicateurs socio-économiques avant de proposer une métrique entre les zones de Chicago, fondée sur cette analyse.

2.1 La criminalité : un phénomène localisé

A chaque crime est associé sa localisation, sous la forme de coordonnées latitude/longitude. Nous avons donc décidé de créer une fonction permettant de représenter les zones de fortes criminalité sur une carte de Chicago. Si l'on applique cette fonction à l'ensemble des crimes commis au cours de l'année 2017, on obtient la représentation donnée en figure 2. Sur cette représentation, il semble se dégager trois zones avec beaucoup de crimes. Une autre information à notre disposition concerne le type de crime commis. On peut représenter la répartition des crimes par type :

Il est frappant de constater que certaines zones concentrent la majorité des crimes d'un type particulier. Ce constat nous invite à chercher des corrélations entre le nombres de crimes commis et des indicateurs

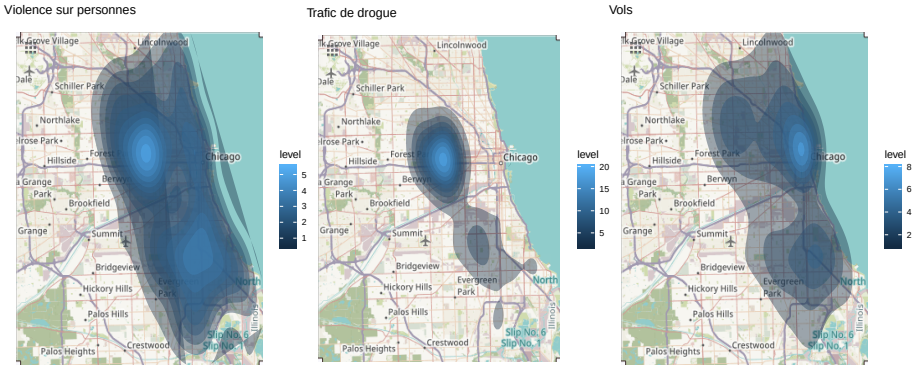

FIGURE 5 – Représentation spatiale des crimes commis en 2017 par type

socio-économiques disponibles pour la ville de Chicago.

2.1.1 Les community areas et les indicateurs associés

Les community areas forment un découpage du territoire de la ville en 77 zones. Leur définition dans les années 1920 par le comité de recherche en sciences sociales de l'université de Chicago avait pour but d'établir des échelles de référence pour les études statistiques.

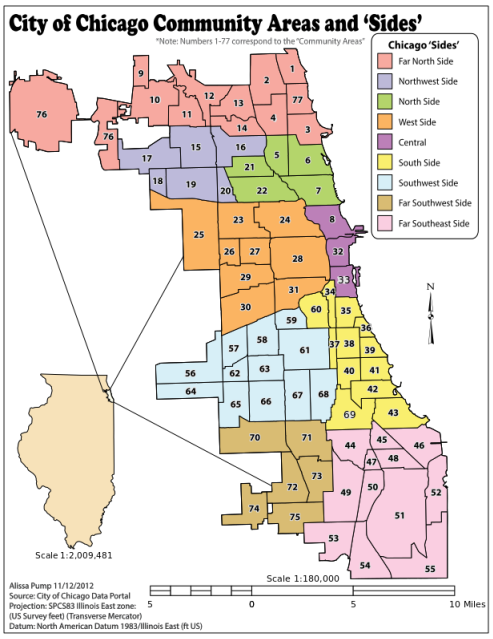

FIGURE 6 – Carte des community areas de Chicago

C'est donc naturellement que les indicateurs socio-économiques issus du recensement sont disponibles par community area. Nous avons accès à différents indicateurs, relevés en 2012, à savoir le taux de chômage, la proportion de la population sous le seuil de pauvreté, le revenu moyen par habitant, le taux de personnes non diplômées de l'enseignement secondaire, le taux de logements insalubres, le pourcentage de la population âgée de moins de 18 ans ou de plus de 64 ans ainsi qu'un indice de précarité. A l'aide d'une analyse en composante principale, nous avons obtenu une synthèse de ces informations.

Le premier axe principal explique 66.3% de la variance observée. Il est négativement corrélé avec le revenu moyen et positivement corrélé avec toutes les autres variables. On interprète facilement cet axe principal comme un indicateur de la richesse de chaque community area.

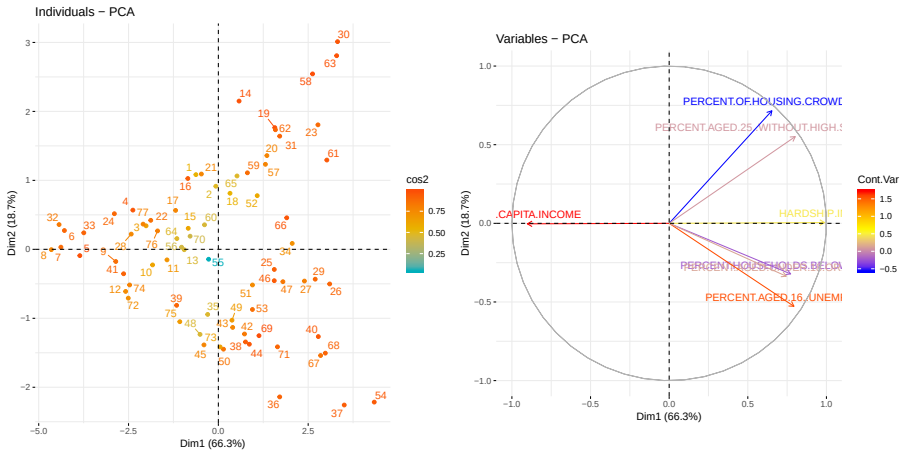

FIGURE 7 – Résultat de l'ACP sur les community areas

2.1.2 Mise en perspective avec les données concernant le crime

Le jeu de données initial contient également, pour chaque crime, la community area dans laquelle il s'est produit. Cette donnée va nous permettre d'étudier l'influence des indicateurs socio-économiques sur la criminalité.

Revenons dans un premier temps à la répartition spatiale des vols et du trafic de drogue. La plupart des vols ont lieu dans la community area Loop et Near North Side. Si on s'en tient aux données de recensement, ce sont deux des quartiers les plus aisés de la ville. En effectuant une recherche sur cette zone, on apprend qu'elle correspond au centre névralgique de Chicago, concentrant bureaux et commerces. Les crimes liés au trafic de drogue se concentrent dans une zone plus à l'ouest vers West Garfield Park, East Garfield Park et North Lawndale. A l'inverse ce sont des quartiers plutôt défavorisés et tristement célèbres pour leur dangerosité, on peut lire par exemple l'article intitulé "West Garfield Park : Chicago's highest homicide rate, lowest life expectancy", publié dans Chicago Tribune le 8 août 2014.

Cherchons maintenant un lien entre le niveau de vie et le taux de criminalité dans chaque community area. Pour ce faire, nous ajustons un modèle linéaire entre le nombre de crimes par habitant dans la community area et le premier axe de l'analyse en composantes principales.

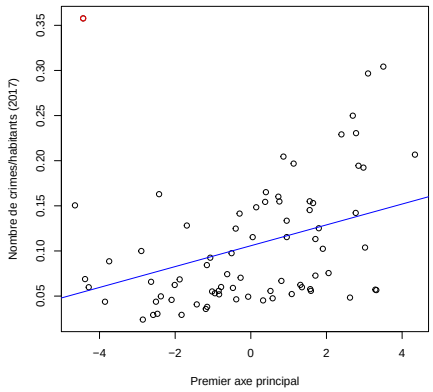

FIGURE 8 – Taux de crimes par habitant en fonction de la richesse

On remarque que le niveau de vie et le taux de criminalité présentent une forte corrélation. Cependant, une community area se présente comme une donnée aberrante, c'est le quartier Loop, en rouge sur le graphique. L'explication, avancée notamment dans Wang *et al.* [12], est que, contrairement aux autres community areas, que l'on pourrait qualifier de résidentielles, l'activité à Loop n'est pas le fait seul des habitants de la zone, mais de l'ensemble des personnes qui y affluent chaque jour pour travailler ou profiter des commerces.

2.1.3 Un indicateur concernant la vie nocturne

En plus des indicateurs socio-économiques caractérisant le niveau de vie des habitants, il nous faut quantifier l'activité de chaque community area. Pour cela, on va utiliser un jeu de données disponible librement sur le site data.cityofchicago.org et recensant les licences délivrées pour l'ouverture de bars du 1er janvier 2006 à aujourd'hui. A partir de ces informations, nous créons une variable correspondant au nombre de licences délivrées par community area.

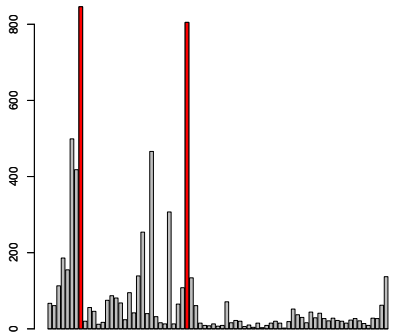

FIGURE 9 – Nombre de licences délivrées depuis 2006 par community area

Cette information nous permet de quantifier l'activité de chaque community area, ce que les données socio-économiques ne permettaient pas. On remarque que dans les community areas numérotées 8 (Near North Side) et 32 (Loop), en rouge sur la graphique, on délivre bien plus de licences qu'ailleurs, ce fait s'explique par le fait que cette zone correspond à la partie la plus active de la ville.

Nous appelons bars la variable centrée réduite associée. L'intérêt de centrer et de réduire est de rendre cette variable comparable aux axes principaux de l'analyse en composantes principales.

2.2 La criminalité un phénomène spatio-temporel

Notre objectif est de montrer que l'évolution de la criminalité n'est pas identique selon les quartiers au cours du temps. Si nous nous intéressons aux deux extrêmes sur le spectre des community area pour la quantité de vols commis, nous obtenons les résultats suivants (en légende le numéro de la community area étudiée). Nous remarquons une augmentation prononcée depuis 2015 pour les community areas Loop (32) et Near North Side (8) du nombre de vols.

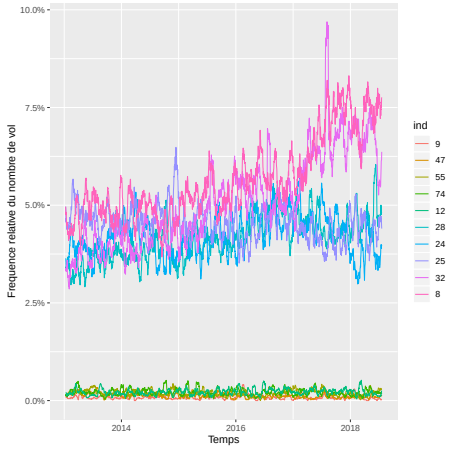

FIGURE 10 – Comparaison de l'évolution du nombre de vols dans le temps entre différents quartiers de Chicago

2.3 Une métrique entre community areas

Pour conclure cette partie, nous présentons une métrique entre community areas que les observations précédentes nous suggèrent. Cette métrique ne correspond pas à une notion de proximité spatiale mais se base sur les critères socio-économiques et l'indicateur concernant la vie nocturne. Nous pensons qu'elle est pertinente du point de vue du problème considéré, au vu des observations précédentes.

Chaque community area est représentée par un point dans $\mathbb{R}^2$, son abscisse correspond à sa valeur pour le premier axe principal de l'ACP, c'est à dire à un indicateur du niveau de vie de ses habitants, et son ordonné à la variable bars, c'est à dire à un indicateur de sa vie festive. On obtient la représentation suivante :

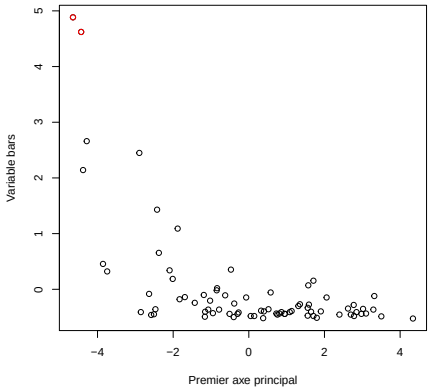

FIGURE 11 – Métrique entre community areas

On apprécie ici le gain d'information obtenu avec l'introduction de la variable bars. En effet c'est elle qui nous permet de distinguer nettement les community areas 8 et 32 (en rouge) des autres, et nous avons également remarqué que la répartition des crimes était bien différente dans ces quartiers qu'ailleurs. Ces observations nous rendent confiants dans le fait que la métrique ainsi construite entre community areas peut être exploitable pour résoudre le problème d'estimation du nombre de crimes.

3 La criminalité : un phénomène temporel multi-dimensionnel

3.1 Quelles sont les variables explicatives d'intérêts ?

Nous avons observé dans l'introduction, comme de nombreux auteurs, que la criminalité est un phénomène intrinsèquement temporel. Nous nous demandons dans cette partie si il ne serait pas pertinent d'introduire un certain nombre de variables exogènes permettant d'expliquer la quantité de crime au cours du temps. Jusqu'ici nous nous sommes intéressés seulement à la composante spatiale de la criminalité. Les variables introduites dans la partie précédentes sont fixes au cours du temps cependant nous voulons prédire la criminalité qui est un phénomène spatio-temporel.

Dans un premier temps, nous avons pensé introduire des données météorologiques comme la température ou la quantité de précipitation. Pour ce faire, nous utilisons les données météorologiques du site National Weather Service [10]. Ces variables météorologiques ont déjà été identifiées comme déterminantes par F.Landau et Daniel Fridman en 1993 [7]. Selon eux, ces variables peuvent être vues comme des proxy pour donner une idée de l'activité des agents au cours de la journée. S'intéresser aux activités des agents est important pour prédire la criminalité car l'action de commettre un crime peut être vu comme un processus proie/prédateur où l'idée de "routine activity approach" est centrale. Ainsi, une certaine température serait liée à un certain comportement des agents prédateurs mais aussi des proies.

Ensuite, nous avons ajouté d'autres variables exogènes comme des variables temporelles pures que sont la place du jour dans l'année et la place du jour dans la série temporelle. Ces dernières sont pertinentes étant donnée l'analyse initiale de la temporalité de la criminalité au cours de l'année.

Dans un deuxième temps, il semblerait vraisemblable que la criminalité au cours du temps soit influencée par des variables socio-économiques comme le taux de chômage ou le taux de pauvreté. Ce constat renforce une remarque déjà faite précédemment. Cependant, il nous a fallu trouver des variables socio-économiques dont nous pouvions avoir la mesure régulière au cours du temps. Par chance, le Bureau of Labor Statistics, équivalent de l'INSEE en France, donne accès à l'évolution mensuelle d'indicateurs comme le taux de chômage ou l'indice des prix à la consommation (CPI) [1]. On rappelle que l'indice des

	types de variables	nom des variables
Y	77 séries spatio-temporelles	nbr de crimes par cmty area
X	2 séries temporelles (mensuelles)	taux de chômage, CPI index
	6 séries temporelles (journalier)	qté de précipitat°, qté de neige, Tmax, Tmin, tdc, jdl
	2 variables dummy	début janvier, fin juillet/début août
	29 variables spatio-temporelle (journalières)	vitesse moyenne par région

prix à la consommation est le prix moyen payé par les citadins pour un panier de biens type au cours du temps. Ces deux variables ont pour objectif de capturer la santé économique de la région de Chicago.

Enfin, une dernière idée qui est inspirée de l'article de Wang *et al.* de 2016 [12] est d'ajouter des variables spatio-temporelles telles que les courses de taxis pour expliquer la criminalité à un certain endroit. Nous aurions pu y avoir accès grâce au portail en ligne de la ville de Chicago cependant la gestion d'une telle base de données aurait demandé des compétences que nous n'avons pas. En ce sens, nous avons pu récupérer un indicateur spatio-temporel qui est la vitesse moyenne des véhicules dans 31 régions de la ville de Chicago. Ces données nous donnent accès en temps réel (toutes les 5 minutes) à la vie de la ville. En effet, si la vitesse moyenne est bien en dessous de la vitesse autorisée alors cela signifie qu'il y a une congestion. Inversement, si le trafic est fluide alors nous nous attendons à ce que la vitesse moyenne soit proche de la vitesse autorisée qui est 50km/h en ville. En bref, nous avons à disposition les variables suivantes :

La base de données que nous utilisons est au pas de temps journalier. Mais nous avons construit une fonction qui nous permet de conserver la base de données originale et d'agréger les données selon un pas de temps et un type de crime voulus. Ceci nous permet de ne pas altérer la base de données initiale et de permettre à d'autres d'explorer certaines interactions non couvertes par le présent rapport. Ainsi, nous nous intéressons à la criminalité par jour de 2015 à 2018. Nous souhaitons maintenant explorer davantage les relations qui existent entre ces différentes variables explicatives et la criminalité à Chicago.

3.2 Exploration des liens entre variables explicatives et criminalité à Chicago

Dans cette partie, nous donnons un premier aperçu des relations entre les différentes variables du jeu de données. Ces premières analyses nous guiderons pour la modélisation future de la criminalité. Les analyses descriptives suivantes sont faites pour le nombre global de crimes dans la ville de Chicago. Le phénomène au cours du temps observé est représenté ci-dessous :

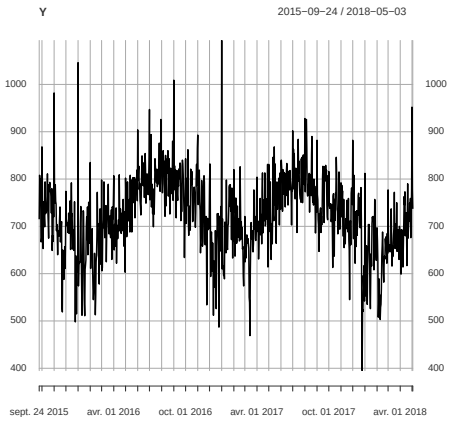

FIGURE 12 – Nombre de crime sur toute la période

D'une part nous constatons que la criminalité est un phénomène rare, en effet, la ville de Chicago compte plus de 2.7 millions d'habitants, et il y a en moyenne seulement 722.32 infractions commises par jour. D'autre part, le phénomène est variable de l'ordre de 81.16 vols par jour. Sur la figure 13, on observe clairement que la période entre Janvier et Mars semble difficile à prévoir quelque soit l'année. Enfin, le nombre de crime est en règle général un processus relativement bruité dans le sens où il y a beaucoup de pics non réguliers. Pour minimiser l'impact de certaines irrégularités notamment en Août et début Janvier, nous avons introduit deux variables 0-1 pour ces périodes où il y a un nombre important de crimes. Nous ne pouvons expliquer ces valeurs aberrantes mais nous pensons judicieux de les prendre en compte dans un modèle séparé.

Nous analysons séparément les vitesses moyennes pour les 31 régions de la ville. Nous utilisons comme précédemment une ACP afin de mieux visualiser les corrélations entre variables. Ici, nous constatons qu'il y a deux groupes dans les variables de vitesses qui apparaissent. Une des conclusions de cette analyse est que les variables de vitesses sont fortement corrélées entre elles.

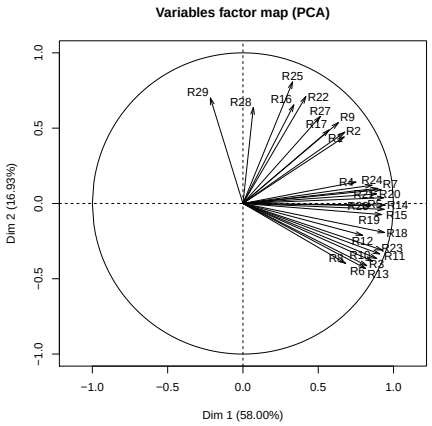

FIGURE 13 – Cercle des corrélations

En analysant plus précisément l'influence de la pluie sur la criminalité, on remarque que le nombre moyen de crimes par jour est significativement plus faible lorsqu'il pleut que lorsqu'il ne pleut pas comme l'indique la p-valeur très faible obtenue par un test de Student de différence de moyenne de l'ordre de 7×10^{-8}. Puis, nous présentons plus en détail sur la relation entre le nombre de crime, la température maximale et la quantité de neige.

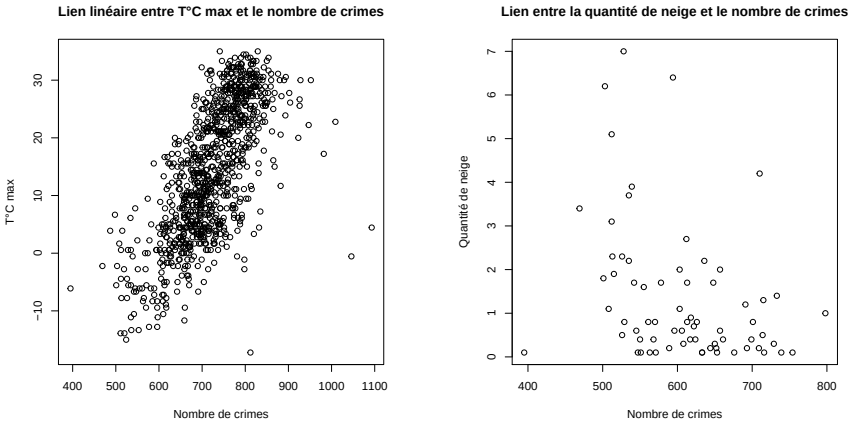

FIGURE 14 – Lien entre la criminalité et les données météorologiques

Nous constatons que plus il fait chaud plus le nombre de crimes est importants. Pouvons nous parler d'un effet chaleur sur la criminalité ? Nous pensons plus à un effet indirect influençant les comportements et donc le nombre d'occasion de commettre un crime. La neige quant à elle, serait corrélée négativement avec le nombre d'infractions. Toutes ces remarques nous guiderons ensuite dans la partie modélisation.

Les résultats obtenus ci-dessus en terme de corrélation le sont pour la totalité des crimes commis à Chicago. Il est important de constater que mise à part la température aucune des variables ajoutées n'est fortement corrélée avec la criminalité. Les variables de température sont fortement corrélées entre elles et les variables de tendance économique le sont avec la variable de tendance. Dans la suite, il nous faudra choisir parmi ces variables pour éviter de trop fortes corrélations au sein des variables explicatives. Les variables TMAX et TMIN étant fortement corrélées, nous choisissons de ne conserver que la variable TMAX. Un point important est que mise à part les corrélations évoquées ci-dessus les co-variables ne sont pas corrélées entre elles. Nous nous intéressons dans la suite à voir si pour certaines échelles spatiales données ces relations entres variables explicatives et variable à expliquer changent. En s'intéressant aux corrélations entre variables, nous faisons déjà un pas vers la compréhension du processus de la criminalité.

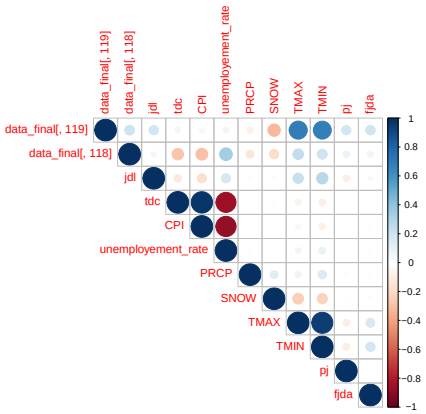

FIGURE 15 – Graphique des corrélations

4 Modélisation de la criminalité à l'échelle globale

La ville de Chicago a besoin d'un outil de modélisation performant pour anticiper les différents endroits à risque à court et moyen terme. Elle s'est déjà équipée de capteurs sonores dans la rue ainsi que de caméras, voir l'article suivant [6], pour réagir plus vite aux différents crimes commis dans la ville. Cette initiative est suivie par une équipe : Strategic Decision Support Centers (SDSCs). Cette dernière est reliée à un laboratoire de recherche Crime Lab associant des chercheurs de l'université de Chicago, qui travaillent sur la problématique de la réduction de la criminalité, à la police. Elle a pour but de déployer les forces de police au bon endroit au bon moment. L'outil de prévision que nous proposons va dans ce sens. Nous nous intéressons aux caractéristiques de la criminalité à Chicago. Dans cette partie nous exposons et critiquons les différents modèles à la fois pour leurs qualités prédictive et d'interprétation. Nous nous appuyons sur des données agrégées à l'échelle journalière allant de début 2015 à début Mai 2018. Notre but est de prédire pendant 30 jours durant le mois d'avril la quantité de crimes commis chaque jour.

4.1 Un problème de prévision

En termes techniques, pour la prévision, nous découpons notre échantillon en deux, un pour l'entrainement et l'autre pour le test. Ainsi nous avons un peu moins de 3 ans de données et nous prenons 1 mois de données pour tester nos modèles et les comparer à la fin de notre analyse. L'ajustement des paramètres se fera à l'aide de la validation croisée par blocs qui est communément utilisée pour la prévision de série temporelle. Nous n'utilisons pas la version plus compliquée qui consisterait à supprimer une partie des données par blocs pour respecter encore davantage l'hypothèse d'indépendance. Nous mesurons l'erreur à l'aide du Root Mean Square Error (RMSE) et du Mean Absolute Percentage Error (MAPE). Ces précisions étant faites, nous pouvons maintenant nous intéresser à la première modélisation. Dans ces présentations de modèles, nous construisons notre boîte à outils pour la dernière partie qui consistera à modéliser à l'échelle locale la criminalité dans les différents quartiers de Chicago.

4.2 Le modèle naïf

La modélisation la plus naïve à laquelle nous avons pensé est de prédire le nombre d'infractions par la moyenne du nombre d'infractions sur la période passée. C'est un des modèles qui sert de référence dans l'article de Wilpen Gorr *et al.* de 2003 [5]. Cette première modélisation revient à écarter toutes influences à la fois de variables exogènes et de la temporalité du phénomène. Une autre un peu plus futée serait de conditionner en fonction du jour dans l'année. Pour plus de détail, nous conseillons d'aller voir l'article de référence.

La conclusion de cette méthode est qu'il y a sur la période considérée en moyenne 722.63 crimes par jours. L'erreur de validation croisée est de 79.55. Enfin, cette méthode fait environ 7% d'erreur pour prédire le mois d'avril. Elle permet bien de capter l'ampleur du phénomène mais non sa variabilité dans le temps. C'est pourquoi nous allons voir une autre méthode de référence couramment utilisée par la police qui est un modèle uni-varié.

4.3 Le modèle univarié : Holt-Winters

Une des méthodes utilisée régulièrement afin de prédire le nombre de crimes à court-moyen terme a trait aux modèles uni-variés. Un bon exemple est encore une fois l'article de Wilpen Gorr *et al.* de 2003 [5]. Il faut bien voir qu'il y a une diversité de modèles uni-variés. Étant donné que ces méthodes ne seront pas le cœur de notre modélisation, nous avons choisi la méthode de lissage exponentiel (Holt-Winters) ajustée automatiquement à l'aide des données et de la fonction `tbast` du package `forecast`. Comme indiqué précédemment la criminalité est un phénomène périodique. On s'attend en ce sens à obtenir de meilleurs résultats à l'aide de la modélisation de cette période.

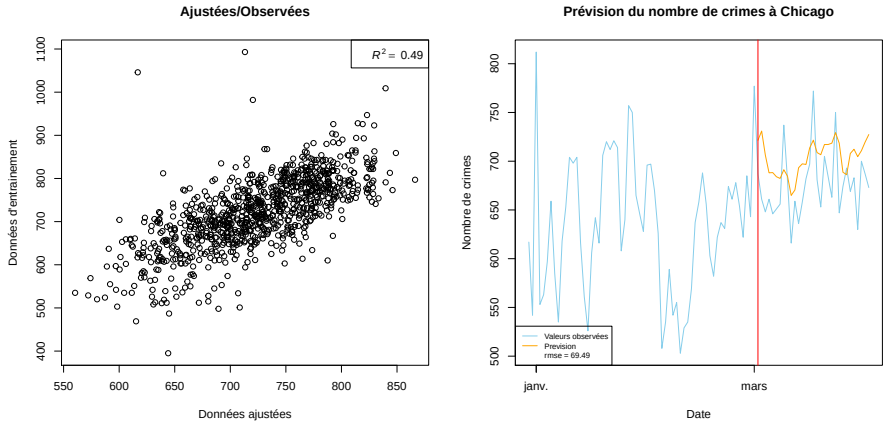

FIGURE 16 – Résultat obtenu avec Holt-Winters

Ce graphique peut nous donner l'impression que la méthode uni-variée n'est pas mauvaise car elle capture parfois le signal sur les données d'entrainements. D'autant que son erreur d'entrainement au carré est de 58.34 comparativement à celle du modèle naïf qui est de 81.56. Par contre pour la prévision à 31 jours ses performances sont très mauvaises, elles sont même pires que la méthode naïve.

Nous pouvons faire une analyse des résidus obtenus par notre modèle. Les différents graphiques de la figure 17 indiquent bien que nous obtenons après ajustement du modèle une erreur qui se rapproche d'un bruit blanc.

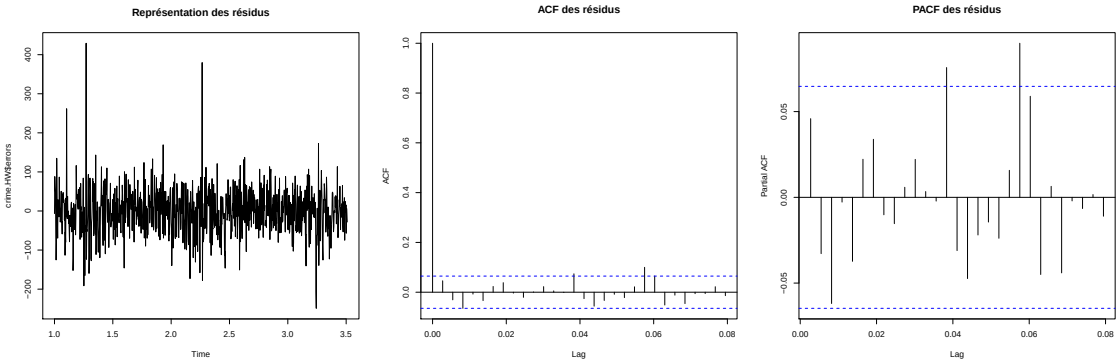

FIGURE 17 – Analyse des résidus pour le modèle Holt-Winters

Cette prévision est la meilleure que nous ayons obtenue avec la méthode de lissage exponentiel. Nous avons tenté une modélisation à l'aide de ARIMA et SARIMA mais sans grands succès. Ce qui nous amène à considérer par la suite des modèles plus complexes au sens où ils prennent en compte des variables exogènes à la quantité de crimes commis. Une des conclusions de cette partie est que le crime n'est pas un phénomène stationnaire ou du moins pas sans traitement statistique de la tendance et de la saisonnalité. L'objectif affiché pour la suite est de faire mieux que les performances des deux modèles précédent : uni-varié et naïf.

4.4 Le modèle linéaire

Nous nous intéressons au modèle linéaire car c'est un modèle simple qui permet d'introduire l'ensemble des variables présentées précédemment. Ce modèle est d'autant plus pertinent qu'il suppose des relations

linéaires entre variable à expliquer et co-variables, ce que nous observons, par exemple entre la température et le nombre de crimes.

Nous constatons que certaines variables sont pertinentes pour modéliser la criminalité comme la température, les variables de précipitations ou le jour dans l'année. En effet, les coefficients associés aux variables jdl, SNOW, PRCP et TMAX sont significatifs au sens du test de Student au risque 0.1 %. Les coefficients pour SNOW et PRCP sont négatifs tandis que celui associé à la température est positif. Ceci signifie que si la quantité de neige augmente alors le nombre de crimes diminue et inversement pour la température. Nous ne pouvons encore une fois faire la conjecture appuyée sur certains articles que le coût de commettre un crime est plus élevé lorsque le temps n'est pas clément.

Nous nous sommes demandés si nous ne pouvions pas introduire une variable retard en supposant que la police veuille prédire le nombre de crimes par jour sur trente jours mais avec une connaissance du nombre de crimes du jour précédent. Il s'avère que nous obtenons une modélisation aux performances de prédiction bien meilleure. Cependant, ce n'est pas l'outil que nous voulons construire. En effet, celui que nous proposons est orienté vers la prévision à moyen terme de la criminalité, non à très court terme. Pour finir de valider notre modèle linéaire, nous faisons une analyse des résidus pour la régression linéaire obtenue ci-dessus.

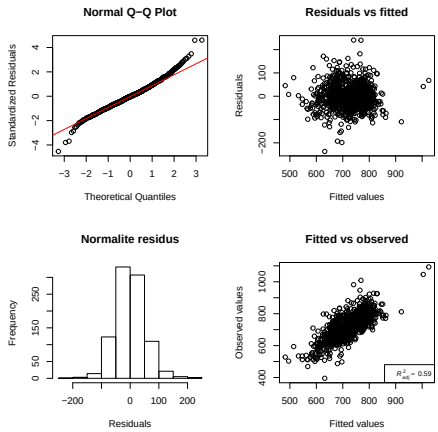

FIGURE 18 – Analyse des résidus du modèle linéaire

Nous constatons que les résidus sont répartis uniformément en fonction des valeurs estimées. De plus, nous obtenons un modèle qui arrive à capter une quantité non négligeable de la variance.

4.5 Le modèle additif généralisé

Nous continuons la présentation des différentes méthodes utilisées pour prédire la quantité de crime sur un mois. Nous adoptons une approche machine learning pour les prochaines méthodes présentées : c'est-à-dire que nous nous focalisons sur l'amélioration des performances prédictives en terme de RMSE et MAPE. Ainsi, nous privilégions les modèles robustes et qui s'ajustent automatiquement. Une méthode largement utilisée pour la prévision de charge à Électricité de France est la méthode gam. Nous la testons sur nos données. C'est une méthode semi-paramétrique capable de capter les irrégularités. Le modèle gam a été développé notamment par Simon Wood dans le package `mgcv`. Nous testons sa performance pour notre jeu de données mais nous sommes relativement sceptiques car nous n'avons pas de variables explicatives qui serait très fortement corrélée de façon non linéaire avec la criminalité à Chicago.

Cependant, le modèle gam est plutôt performant pour notre modèle si on en croit l'erreur d'entrainement et l'erreur de validation croisée. D'autant que sur la figure 20, on remarque qu'il capture une partie de la variabilité du phénomène et qu'il laisse une partie de ce dernier sans modélisation. La question est de savoir si les pics erratiques sur la série sont dus à du bruit ou à des variables explicatives manquantes. L'hypothèse d'un bruit blanc n'est pas à rejeter si nous pensons au corrélogramme présenté précédemment pour le modèle Holt-Winter.

4.6 Les forêts aléatoires

Les forêts aléatoires reposent sur la méthode des arbres CART qui sont des arbres de décision. Ces arbres sont très non paramétriques et ils sont aussi peu sujets au sur-apprentissage. Les méthodes d'arbres sont beaucoup utilisées dans les méthodes d'ensemble comme le boosting. Ainsi, nous les utiliserons dans

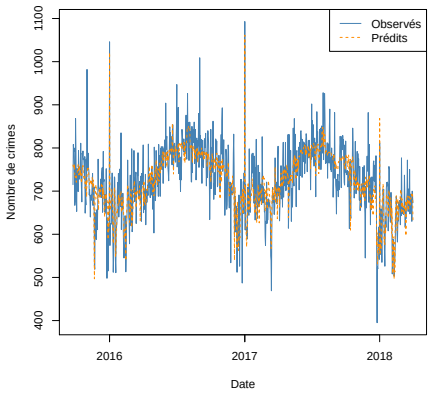

FIGURE 19 – Prévision pour le modèle gam

la partie agrégation d'experts bien qu'ils ne soient pas parmi les meilleurs prédicteurs. Nous utilisons le package `RandomForest`.

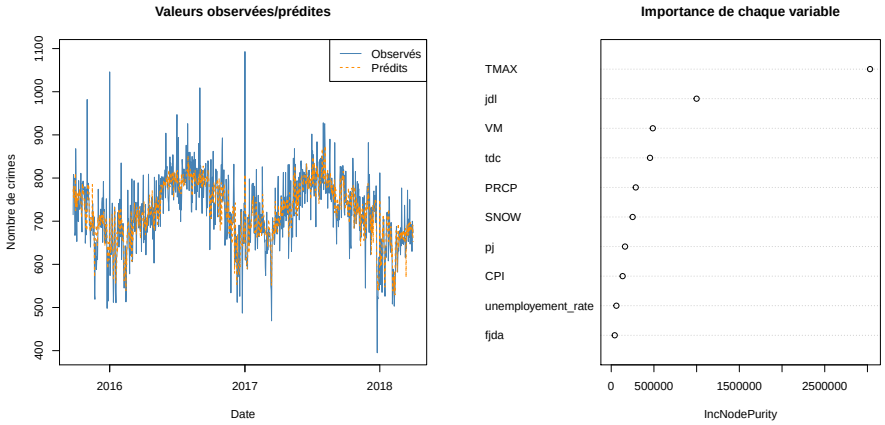

FIGURE 20 – Résultats de la forêt aléatoire

Par rapport au modèle gam, la forêt aléatoire permet de capter plus de variabilité sur les données d'entrainements. C'est un modèle intéressant pour la suite. D'autant que nous pouvons avoir accès à la mesure de l'importance qui permet de classer grâce à l'erreur outofbags les variables par leur influence dans la construction de l'arbre. Ainsi, nous avons la confirmation que la vitesse moyenne peut avoir un intérêt.

4.7 La modélisation en grande dimension

Nous sommes amenés à nous demander si introduire des interactions ou des ordres supérieurs des variables explicatives ne pourraient améliorer les modèles. Nous générons pour ce faire toutes les variables croisées ainsi que les carrés de ces variables à l'aide de la fonction `poly` de R. L'augmentation de la dimension de notre problème entraine une plus grande flexibilité des modèles. Nous utilisons un algorithme de sélection de variables type forward sélection.

Un des problèmes de l'augmentation de la dimension utilisée ci-dessus est son manque d'interprétabilité. Mais nous constatons tout de même que la température est toujours une variable déterminante. La variable croisée taux de chômage et tendance ressort aussi comme une variable corrélée au nombre de crimes. Ceci peut s'expliquer par l'influence de la santé économique de la région de Chicago sur la criminalité. En effet, on constate que sur la période considérée entre 2015 à 2018, le chômage diminue. On peut se demander si il ne reste pas de la corrélation temporelle dans les résidus pour cela nous comparons les prévisions du modèle avec et sans utiliser un processus ARIMA pour les résidus.

En conclusion du modèle linéaire, il ne parait pas pertinent de modéliser les résidus à l'aide d'un modèle ARIMA. L'augmentation du nombre de variables nous invite à utiliser des méthodes comme la régression ridge ou le LASSO pour modéliser le nombre de crimes.

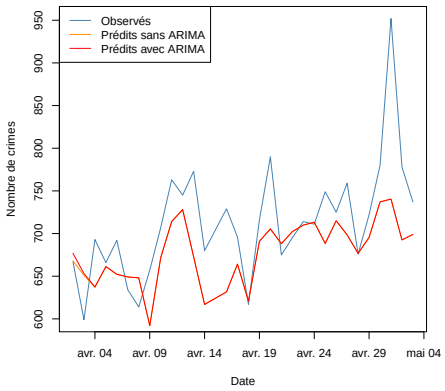

FIGURE 21 – Comparaison de la prévision pour un modèle linéaire avec et sans ARIMA

Nous pouvons donner une analyse succincte des résultats obtenus en terme de variables sélectionnées pour la méthode LASSO. Nous constatons que le nombre de variables sélectionnées est 13. Il y a finalement un faible nombre de variables sélectionnées. Si nous nous intéressons de près à ces dernières ce sont principalement des variables croisées liées à la température ou à la quantité de neige tombée (nous sommes au nord des États-Unis). Ce nombre est proche de celui sélectionné par la méthode forward pour le modèle linéaire. Ces variables sélectionnées sont comme attendu liées d'une manière ou d'une autre aux variables PRCP, TMAX, SNOW ou le jour dans l'année.

4.8 Le boosting

Enfin, la dernière méthode prometteuse est le boosting. Nous allons l'implémenter à l'aide de deux package, `gbm` et `xgboost`. Le boosting en théorie apprend sur des distributions différentes des données d'entrainements au cours de l'apprentissage. La méthode employée peut se rapprocher de l'apprentissage par renforcement d'une certaine manière. Nous présentons les résultats de ces méthodes dans le tableau bilan ci-dessous.

4.9 Bilan de la prévision à l'échelle globale

Le but de cette partie est de faire un bilan général sur les méthodes présentées. On ne peut pas s'empêcher ici de faire un catalogue des résultats à la fois en terme d'erreur de validation croisée et d'erreur apprentissage/test.

	Méthodes	RMSE VC	Ecart type RMSE VC	RMSE : test-train	Mape : test-train %
1	naif	79.55	23.28	66.53	6.85
2	HW	87.48	25.36	69.49	6.96
3	lm	53.41	8.77	46.54	4.78
4	lm.select.fwd	51.67	7.62	60.38	5.77
5	gam	52.10	9.26	61.13	5.89
6	rf	53.30	9.92	51.87	5.09
7	lasso	57.28	14.95	47.64	4.66
8	ridge	65.86	19.14	48.22	4.77
9	gbm	55.00	12.58	58.20	6.15
10	xgboost	55.00	12.58	53.19	5.47

TABLE 1 – Tableau récapitulatif des erreurs au carré de cross-validation pour les différentes méthodes

Le mois d'avril est un mois facile à prévoir car plus proche de la moyenne des autres années, en effet le modèle naïf fait a une meilleur erreur sur ce mois que pour l'erreur de validation croisée. Ceci explique la meilleure performance de toutes les méthodes en prévision par rapport à leur performance pour l'erreur de validation croisée. La conclusion de ce tableau est qu'il n'y a pas vraiment de méthodes statistiques pour traiter ce jeu de données qui ressortent du lot. Nous allons confirmer cela à l'aide du graphique suivant comparant les prévisions des différentes méthodes sur la partie test.

Quelque soit la méthode nous n'arrivons pas à prédire les principaux pics. Chaque méthode arrive cependant à prédire des choses différentes. En un sens, nous pouvons les voir comme complémentaires.

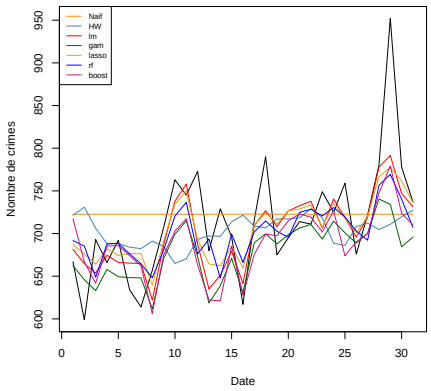

FIGURE 22 – Comparaison des prévisions pour la criminalité globale sur le jeu de données test

Les méthodes de prévision considérées sont toutes meilleures que la méthode uni-variée et que la méthode naïve sur la partie considérée. Nous sommes confiants quant à la performance globale de nos modèles. Nous remarquons tout de même que le modèle linéaire est un bon modèle candidat pour la suite aussi bien en terme d'erreur de validation croisée que d'erreur train/test. De plus, nous faisons au mieux 4.78% d'erreur sur un mois facile mais pour la prévision de crimes cela semble satisfaisant.

5 Agrégation spatiale des modèles

Dans cette partie, nous présentons et nous mettons en œuvre un algorithme d'agrégation spatiale à partir d'un des modèles prédictifs introduits dans la partie précédente et nous analysons le résultat obtenu.

5.1 Un algorithme d'agrégation spatiale

Nous présentons une stratégie d'estimation du nombre de crimes à Chicago qui utilise la métrique entre les community areas définie dans la première partie.

Dans la partie précédente, nous avons construit des modèles prédictifs pour la série temporelle qui correspond à l'ensemble des crimes à Chicago. Cependant, cette approche globale ne tient pas compte des spécificités de chaque community area. Or, nous pensons qu'il pourrait exister des phénomènes plus locaux dans la criminalité à Chicago. Dans cette optique on peut construire un modèle pour chacune des 77 community areas et additionner les prédictions locales pour obtenir une prédiction globale. Entre ces deux approches on peut vouloir faire des modèles intermédiaires en construisant des prédicteurs sur des groupes de community area.

Pour être exhaustif, nous devrions construire autant de modèles qu'il y a de sous-ensemble non vides d'un ensemble à 77 éléments, soit $2^{77} - 1$. Il est évidemment exclu de parcourir tous ces modèles. Nous devons donc chercher une manière raisonnable de parcourir les échelles intermédiaires.

Afin de rendre l'exploration des différentes échelles possibles en temps raisonnable, on utilise la métrique entre community area définie précédemment avec une classification hiérarchique ascendante, on obtient alors un modèle par niveau d'agrégation (c'est à dire par nombre de groupes de community areas constitué).

La stratégie d'estimation est la suivante. On commence par construire 77 modèles, un par community area. Puis, pour chaque regroupement dans la classification hiérarchique, on calcule un modèle pour le groupe formé. L'avantage est qu'à chaque regroupement, on ne construit qu'un nouveau modèle. Ainsi on parcourt les différentes échelles envisageables en ne calculant que 77+76=153 prédicteurs. L'algorithme est le suivant :

5.2 Résultat pour le modèle linéaire

Cet algorithme peut être utilisé avec n'importe lequel des modèles présentés. Dans un premier temps, nous nous focalisons sur l'interprétabilité des résultats et la description des effets des différentes variables introduites. La recherche de performance de prédiction pure fera l'objet d'une étude ultérieure. Dans cette optique, nous choisissons le modèle linéaire sans introduire de variables croisées. La figure 23 indique le RMSE obtenu pour chaque niveau d'agrégation.

Algorithm 1 Agrégation spatiale de modèles

Require: Un modèle prédictif

for $i = 1, \ldots, 77$ **do**
Construire un modèle prédictif pour la community area i
end for
Sommer les prédictions locales pour construire une prédiction globale

for $j = 76, \ldots, 1$ **do**
Concaténer les deux groupes dont les centres sont les plus proches au sens de la métrique sur les community area
Construire un modèle prédictif pour le groupe ainsi formé
Sommer les prédictions dans chaque groupe pour obtenir une prédiction globale
end for

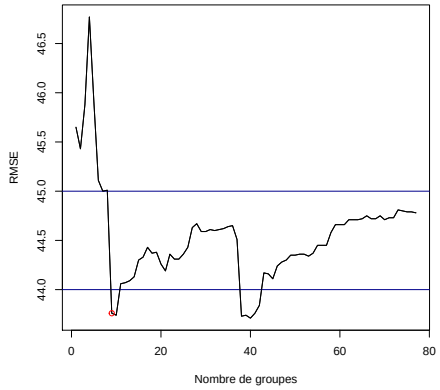

FIGURE 23 – RMSE en fonction du niveau d'agrégation

En suivant la courbe de gauche à droite, on observe des phénomènes intéressants. Pour un faible nombre de groupe (de 1 à 8), le RMSE est le plus mauvais obtenu, oscillant entre 45 et 46.77. Ce constat nous conforte dans l'idée qu'il est judicieux de créer des modèles à différentes échelles afin de prendre en compte dans la prévision les particularités des différentes community areas. Pour un nombre de groupe égal à 9 ou 10, la prévision est proche de la meilleure obtenue. On obtient un RMSE de 43.76 pour 9 groupes, de 43.74 pour 10 groupes lorsque le meilleur score obtenu est 43.71. Ensuite, si l'on considère entre 11 et 37 groupes, la prévision perd en qualité. Le RMSE moyen sur cet intervalle est de 44.41. Puis, entre 38 et 41 groupes, on obtient le meilleur score de prévision, le RMSE étant minimal pour 40 groupes. Le RMSE moyen sur cet intervalle est de 43.73. Enfin, pour 42 modèles ou plus, le RMSE remonte et est compris entre 44 et 45. On observe également que c'est dans ce régime que la qualité prévision paraît la moins instable par rapport au nombre de groupes. Le RMSE moyen sur cet intervalle est de 44.53. L'interprétation que l'on peut en tirer est qu'il n'est pas judicieux d'adopter le point de vue extrême consistant à construire un modèle par community area, mais qu'il est préférable d'en regrouper certaines, semblables du point de vue de l'objectif de prévision.

Cette analyse nous apprend donc que pour obtenir une bonne prévision du nombre de crimes en sommant des prévisions locales, il faut faire un compromis entre la situation où l'on ne considère pas assez de groupes et celle où l'on en considère trop. Cet équilibre peut s'interpréter comme un compromis sous-apprentissage/sur-apprentissage. Ici, il semble exister deux régimes optimaux, le premier autour de 10 groupes et le second autour de 40.

Nous avons donc confirmé la pertinence de notre algorithme multi-échelles pour prédire le nombre de crimes à Chicago ainsi que la métrique utilisée pour former les groupes de community areas. Nous cherchons maintenant à analyser qualitativement les modèles linéaires obtenus. Pour cela nous nous intéressons aux modèles obtenus pour un découpage en 9 groupes, car parmi les niveaux d'agrégation menant aux meilleurs scores de prévision c'est celui qui comporte le moins de groupes. Ce choix est motivé par un soucis de simplicité. En effet, plus on considère de groupes, plus il y a de modèles à décrire nous ne voulons dans ce rapport donner que quelques exemples. Si la ville de Chicago voulait utiliser nos

résultats, elle pourrait reprendre cette étude pour des niveaux d'agrégation et des groupes qui lui serait d'un intérêt particulier.

5.3 Description de certains modèles locaux

Nous nous focalisons donc dans cette sous-partie sur le résultat obtenu pour 9 groupes de community areas. Parmi eux, nous avons choisi ici d'en analyser deux plus finement, car ils présentent un intérêt particulier pour la compréhension du phénomène et font écho avec des remarques déjà formulées. Nous représentons les deux groupes d'intérêt en couleur sur la figure 24 correspondant à la métrique entre community areas.

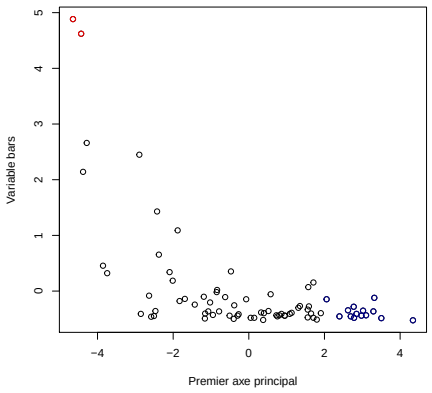

FIGURE 24 – Deux groupes de community areas significatifs

5.3.1 Le cas de Loop et de Near North Side

Le groupe en rouge correspond à deux community areas déjà évoquées précédemment, Loop et Near North Side. Ce sont des quartiers parmi les plus riches de la ville et ils représentent 4% de la population totale, tout en concentrant environ 8% des crimes commis chaque jour. Nous avions déjà expliqué l'intérêt de les considérer séparément, en raison de leur vie festive prononcée. Les résultats obtenus précédemment viennent conforter notre intuition, pour une meilleure prévision, il peut être judicieux de leur allouer un modèle spécifique. Passons maintenant à l'étude détaillée du modèle obtenu. Pour pouvoir comparer les coefficients de la régression linéaire, nous avons centré et réduit les variables. Le tableau suivant donne les coefficients ainsi que les p-valeurs du test de Student pour les variables les plus influentes du modèle.

	coefficient	p-valeur
(Intercept)	57.7	$< 2 \times 10^{-16}$
jdl	1.17	1.01×10^{-2}
tdc	11.00	5.71×10^{-9}
CPI	-3.96	4.81×10^{-2}
unemployement_rate	1.85	3.3×10^{-2}
PRCP	-1.14	8.60×10^{-3}
TMAX	3.56	1.77×10^{-14}

TABLE 2 – Coefficients et p-valeurs du test de Student pour les 7 variables les plus influentes (Loop et Near North Side)

L'intercepte correspond au nombre moyen de crimes commis par jour dans ce groupe de community areas. Hors intercepte, les variables les plus influentes (au sens de la p-valeur du test de Student) sont tdc, TMAX et PRCP.

La variable la plus influente est la température maximale dans la journée (TMAX), et elle est positivement corrélée à la variable de sortie. Au contraire, la variable précipitation (PRCP) est négativement corrélée au nombre de crimes. Cette observation nous permet de retrouver au niveau local une remarque déjà faite lors de l'analyse au niveau global, à savoir que plus la météo est clémente, plus il y a de crimes. Voilà une information qui pourrait s'avérer précieuse pour la ville de Chicago si elle cherche à optimiser la présence de ses forces de l'ordre au cours de l'année.

L'autre variable influente est la tendance (tdc). C'est une variable introduite pour la prévision de série temporelle qui indique le nombre de jours écoulé depuis la date à laquelle débute la série des données utilisées pour ajuster le modèle. Ici cette variable est positivement corrélée au nombre de crimes. Ce résultat est négatif pour la ville de Chicago car il signifie que la criminalité est en hausse pour ce groupe de community areas. Si des politiques pour lutter contre le crime sont mises en place dans cette zone, il serait peut être bon d'en faire la critique et sinon il serait peut être judicieux d'en proposer.

5.3.2 Le cas des quartiers pauvres

Le groupe représenté en bleu correspond aux community areas les plus pauvres de la ville, on y retrouve notamment West Garfield Park, considéré par certains journalistes comme la zone la plus dangereuse de Chicago. Elles regroupent 16% de la population de Chicago, et environ 23% des crimes déclarés chaque jour. Ici aussi nous analysons les variables les plus influentes.

	coefficient	p-valeur
(Intercept)	150.84	$< 2 \times 10^{-16}$
jdl	-1.64	1.08×10^{-2}
tdc	-12.06	1.97×10^{-5}
CPI	8.84	2.05×10^{-3}
PRCP	-2.26	2.20×10^{-4}
SNOW	-2.30	2.21×10^{-4}
TMAX	12.71	$< 2 \times 10^{-16}$

TABLE 3 – Coefficients et p-valeurs du test de Student pour les 7 variables les plus influentes (quartiers pauvres)

Encore une fois, outre l'intercepte, les variables d'influence mises en lumière par le modèle sont les variables liées à la météo (précipitation, température maximale et quantité de neige). On note une corrélation positive pour la température maximale et négative pour les variables précipitation et neige. Dans cette zone également il y a plus de crime par beau temps.

Ici la tendance est également très prononcée, mais cette fois-ci corrélée négativement avec le nombre de crimes observés. Ce constat paraît de prime abord positif quand aux politiques de lutte contre le crime existantes et la ville de Chicago semble avoir tout intérêt à poursuivre dans cette voie. Il faut cependant émettre des réserves sur ces chiffres. En effet, comme nous l'avons déjà remarqué, le nombre de crimes présent dans le jeu de données concerne uniquement les crimes qui ont été signalés. Cette baisse observée peut s'expliquer soit par une baisse réelle de la criminalité, soit par une moins bonne efficacité des effectifs de police. Cependant, les données à notre disposition ne nous permettent pas de trancher.

5.4 Bilan de l'agrégation spatiale de modèles appliquée au modèle linéaire

Ces résultats nous ont permis de confirmer l'intérêt de créer des modèles locaux pour différents groupes de community areas et de sommer les prévisions obtenues pour obtenir une prévision globale. On a également constaté que la métrique utilisée pour agréger les différents groupes était pertinente pour répondre à l'objectif de prévision. Le fait de s'intéresser en particulier à des modèles linéaire nous a permis d'analyser qualitativement les modèles obtenus. Nous avions déjà mentionné l'importance de la météo, nous l'avons ici retrouvée dans chacun de nos modèles. De plus, ces modèles nous ont donné accès à la tendance de l'évolution du nombre de crimes dans certaines zones. Nous avons choisi de nous concentrer sur deux cas significatifs. Cette analyse reste de petite ampleur au vu de l'information disponible. Nous possédons en effet un modèle linéaire pour chaque groupe de community area à chaque niveau d'agrégation. Cette masse d'information ne demande qu'à être utilisée pour étudier qualitativement l'influence relative de différentes variables sur le nombre de crimes dans différents zones de la ville, selon les besoins de la ville de Chicago.

Cette analyse est satisfaisante car elle fournit des sorties facilement interprétables et une prévision correcte, mais en ce sens elle ne répond qu'à la moitié de la problématique initiale. Il nous reste maintenant à voir comment créer des modèles prédictifs plus performants, quitte à perdre en interprétabilité.

6 Agrégation d'experts

Un des objectifs de cette partie est de compléter les analyses faites au-dessus. Notre projet a notamment pour but de proposer une méthode de modélisation capable de prédire de façon précise pour chaque community area le nombre d'infractions commises. Dans la partie précédente nous nous sommes

concentrés sur l'interprétation des résultats obtenus avec le modèle linéaire. Dans cette partie, nous allons présenter une méthode pour convertir des prédictions faites à une échelle vers une autre échelle à l'aide des données passées. Ensuite la méthode d'agrégation spatiale peut être améliorée en considérant une agrégation séquentielle d'experts plutôt qu'une simple somme. Enfin, nous montrons l'intérêt des différentes modélisations suivant la granularité pour augmenter la robustesse de notre méthode de prédiction.

6.1 Changement d'échelle : outils et performances

L'objectif du projet est de prédire la criminalité à la fois à l'échelle locale et à l'échelle globale pour la ville de Chicago. Nous avons eu l'idée de passer des prévisions locales à la prévision globale à l'aide d'un coefficient de proportionnalité appris sur les données passées pour chaque groupe de community areas. La question est de savoir comment techniquement extrapoler les prédictions faites pour un quartier à l'échelle globale ou inversement ? Ces coefficients nous servent à deux reprises : pour faire de l'agrégation d'experts et pour comparer les performances des modèles locaux par rapport à une désagrégation spatiale de la prédiction globale. Ainsi, nous évaluons l'intérêt d'avoir des modèles spécifiques à chaque community area contre un modèle global que nous désagrégeons. En effet, pour la police de Chicago, il semble plus simple d'avoir un outil général pour l'ensemble de Chicago qui ne dépend pas de données locales comme la vitesse des véhicules. Cela est d'autant plus vrai qu'un modèle global désagrégé leur éviterait d'avoir à estimer un grand nombre de paramètres pour leurs différents modèles. Cependant, nous montrons qu'il est souhaitable pour la police de Chicago d'avoir un outil qui prenne en compte les spécificités de chaque quartier.

En effet, si on considère les résultats suivants pour la métrique calculée suivant la formule ci-dessous. Soit K le nombre de cluster, nous appliquons la formule suivante pour comparer les deux stratégies locale et désagrégée.

$$\sum_{k=1}^{K} RMSE^k_{desagrege} - RMSE^k_{local}$$

Nous constatons que quelque soit l'échelle choisie la différence définie ci-dessus est positive. Cela est vrai pour une granularité à 9 groupes dont la différence est positive de l'ordre de 5. Ce qui est aussi vérifié pour le cas où on prend chaque community area individuellement.

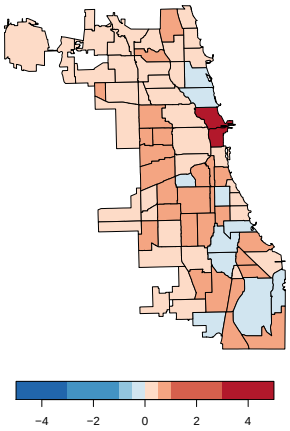

FIGURE 25 – Différence d'erreur entre désagrégation spatiale et modélisation locale

En étudiant plus précisément ces différences on constate que la différence d'erreur est la plus importante à Loop et Near North Side. Pourtant ces quartiers ne concentrent pas la majeure partie de la criminalité qui est localisée dans les quartiers pauvres. Cependant, réussir à capter les variations de ces deux community areas est un facteur déterminant pour améliorer nos prévisions. Ce qui peut s'expliquer par les raisons déjà évoquées.

6.2 Performances de quatre modèles pour les différents clusters

Dans cette sous partie, nous étudions les performances en terme de prévision de quatre modèles : le modèle gam, le modèle linéaire, le modèle linéaire avec sélection de variables et le LASSO. L'objectif est de voir si nous arrivons à retrouver des schémas pour certaines échelles concernant les performances de prévision. Dans le cas présent, nous tenterons d'expliquer ces variations entre méthodes.

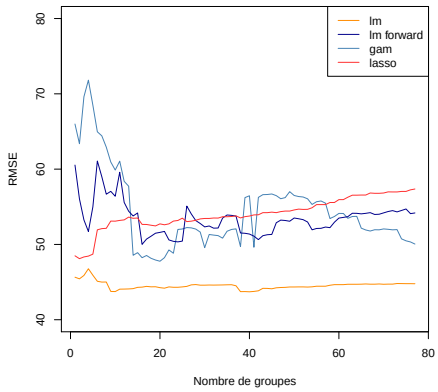

FIGURE 26 – Différents modèles pour différentes échelles spatiales

On constate que l'erreur de prévision diminue pour trois des quatre méthodes lorsque l'on désagrège un peu et a tendance à remonter lorsque nous désagrégeons trop les groupes pour obtenir des modèles de petites tailles. Ceci peut s'expliquer par un compromis sur-apprentissage contre sous-apprentissage, c'est-à-dire que les modèles locaux sont meilleurs qu'un modèle global lorsque l'on isole des quartiers qui ont des spécificités particulière (population, activité économique) mais lorsque nous désagrégeons davantage il y a des quartiers qui deviennent difficiles à prévoir car isolés. C'est une des explications déjà avancée dans la partie précédente. Pour argumenter davantage en ce sens, nous allons faire deux choses : regarder si l'erreur de validation croisée suit la même trajectoire pour le modèle linéaire et étudier plus précisément où l'erreur est commise lorsque l'on désagrégé les modèles.

6.3 Étude de l'erreur de validation croisée pour le modèle linéaire

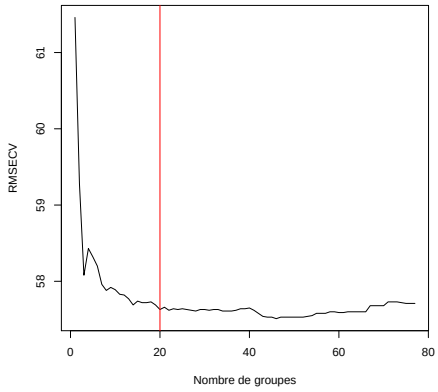

FIGURE 27 – Erreur de validation croisée pour le modèle linéaire

En s'intéressant à l'erreur de validation croisée pour le modèle linéaire. On constate d'une part que le problème de prédiction au mois d'avril est plus simple car l'erreur est plus faible. D'autre part, on constate que le minimum est aussi vers 40-45 groupes. Cependant si on applique une règle du coude sur l'erreur de validation croisée, on aurait tendance à choisir 20 groupes. En ce sens, nous ne retrouvons pas exactement les mêmes résultats que pour l'erreur train/test mais nous avons globalement les mêmes conclusions, c'est-à-dire que nous faisons mieux en ayant des modèles locaux. En appliquant le rasoir d'Ockham, nous choisissons un modèle parcimonieux à 20 groupes par exemple.

Nous faisons maintenant l'étude plus précise de ce qu'il se passe du passage de 4 à 20 groupes en terme d'erreur. Quels sont les quartiers qui influencent le plus la différence d'erreur ? Une des raisons qui explique les meilleurs performances de modèles locaux pour à la fois prédire le global et le local est la capacité des modèles plus petits à sur et sous prédire par rapport à la cible. Leur somme a davantage de chance d'être proche de la cible car la diversité est plus grande. En effet, si on compare la différence moyenne entre prédiction et test pour 4 modèles et 20 modèles, on constate que la différence est moins

négative pour 20 que pour 4.

6.4 Agrégation séquentielle ou somme des prévisions

Dans cette partie, nous allons comparer les performances de deux types d'agrégation différentes en terme d'erreur apprentissage/test pour le modèle linéaire. En effet, dans l'algorithme précédent nous avons décidé de faire la somme des prédictions. Ceci est naturel mais nous avons une autre idée qui est de passer à l'échelle globale chaque prédiction par groupe de community areas pour ensuite les agréger suivant une combinaison convexe apprise séquentiellement grâce au package opera. Nous constatons que cette stratégie améliore l'erreur train/test pour le modèle linéaire. Cette démarche est justifiée par le fait que la ville de Chicago met à disposition chaque jour les crimes de la veille. C'est un schéma de diffusion de donnée qui se prête bien à l'apprentissage séquentiel.

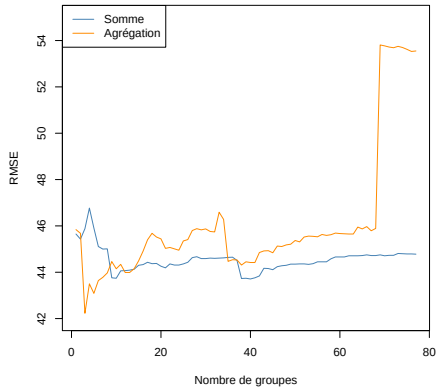

FIGURE 28 – Comparaison entre deux méthodes différentes d'agrégation : une reposant sur la somme des prédictions individuelles, l'autre sur la mise à l'échelle puis une combinaison convexe séquentielle des prévisions

On constate que le fait d'agréger de façon intelligente les prédictions tout en les renormalisants (aux erreurs d'estimation près) permet d'obtenir un modèle désagrégé avec environ 4 groupes qui est très performant. Ainsi, nous allons regarder plus précisément ce qu'il se passe pour l'échelle à 9 groupes et à 4 groupes.

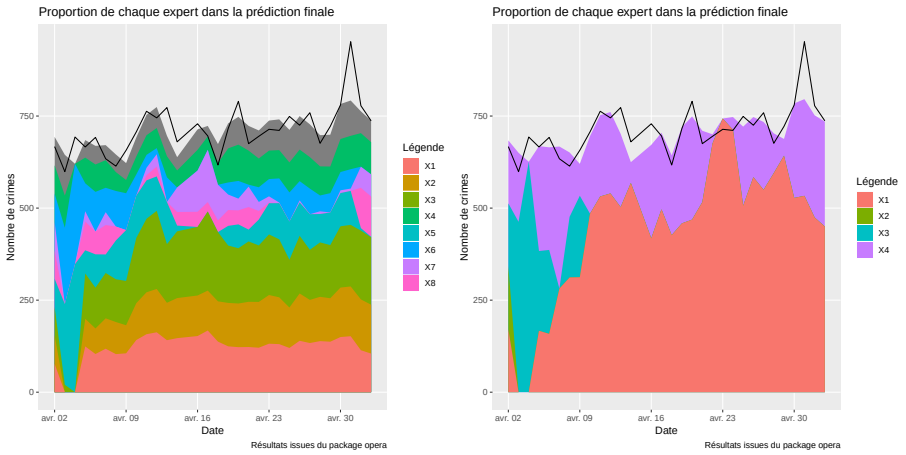

FIGURE 29 – Agrégation d'experts pour différentes échelles pour les prévisions obtenues avec le modèle linéaire

Nous constatons que pour l'échelle spatiale avec 4 groupes, 2 experts ressortent comme principaux lors de l'agrégation. Nous faisons mieux que l'agrégation uniforme potentielle. La température nous permet de prendre en compte l'information temporelle proche dans le passé et ainsi avoir un horizon de prévision de 30 jours comparable à ce que l'on peut obtenir en agrégeant séquentiellement. Pour l'agrégation de 9 experts nous conservons tous les experts. Ces derniers ont tendance à prédire plus de crime sur la période considéré comparativement aux 2 experts pour 4 groupes.

6.5 Agrégation d'experts à l'échelle globale

Enfin, maintenant que nous avons un ensemble d'experts potentiels pour des niveaux de granularité différents, nous utilisons à nouveau une méthode d'agrégation convexe pour améliorer nos performances de prédictions en agrégeant des méthodes dont on veut capter les différents avantages. Nous avons 4 méthodes avec des prédictions à des échelles différentes. Nous choisissons 5 experts par méthode auxquels nous ajoutons les différentes méthodes entrainées au préalable pour prédire le crime globalement. Nous constatons que nous ne faisons pas mieux que le meilleur expert sur la partie à prévoir considérée. Cependant, nous pouvons espérer que l'agrégation convexe de l'ensemble des méthodes permettent d'obtenir une plus grande robustesse de la prévision. Les experts mis en avant sont comme attendu les modèles additif généralisé et linéaire à des échelles favorables.

7 Conclusion

7.1 Réponse à la problématique

Notre objectif était double : prédire efficacement le nombre de crimes par jour et comprendre quels facteurs exogènes jouent sur la criminalité. L'évolution de la qualité de notre prédiction est résumée dans la tableau suivant :

Méthode	RMSE
Modélisation naïve (partie 4.3)	66.53
Modèle linéaire à l'échelle globale (partie 4.4)	46.54
Agrégation de modèles linéaires locaux en sommant les prédictions (partie 5)	43.71
Agrégation d'experts à partir de modèles locaux remis à l'échelle globale (partie 6)	42.23

Table 4 – Évolution du RMSE sur l'échantillon test

Nous voulons mettre en avant le gain obtenu en utilisant l'algorithme d'agrégation de modèles locaux (parties 5 et 6). Cet algorithme réalise une synthèse de l'analyse qualitative de la partie 2 et des méthodes générales d'estimation présentées en partie 4, utilisées avec les variables que nous avons ajoutées au jeu de données (partie 3).

Cette construction nous aura permis de mieux comprendre les facteurs influents la criminalité à Chicago. Comme de nombreux auteurs, nous avons remarqué une influence forte de la météo. De plus nous avons mis en avant la nécessité de traiter dans des modèles séparés certaines zones de la ville et nous avons proposé une métrique entre community areas permettant de définir ces zones.

Au final, nous avons donc proposé un outil performant de la prédiction journalière des crimes ainsi que de nombreuses analyses permettant de mieux comprendre le phénomène.

Nous avons du faire face à quelques difficultés. Premièrement, nous insistons sur le fait que nous n'avons à disposition que les crimes déclarés, et qu'il est possible qu'il soit commis à Chicago de nombreux crimes non présents dans notre base de donnée. Deuxièmement, la police de Chicago bénéficie sûrement d'informations complémentaires qui pourraient être pertinentes pour prédire le crime, mais qui sont confidentielles (un fichier recensant les gangs actifs par exemple). Nous avons fait face à ses obstacles en nous efforçant de rester honnêtes quant à la portée des conclusions que l'on pouvait tirer de notre travail.

7.2 Enseignements

Ce projet aura été l'occasion pour nous de nous confronter à un jeu de données réel et de saisir l'intérêt de chercher à comprendre le phénomène sous-jacent dans sa globalité. Pour ce faire, nous avons étudié la littérature concernant la prévision du crime et d'une certaine manière, reproduit les expériences qui y sont présentées.

Le jeu de données final que nous avons utilisé est en réalité une concaténation de nombreux jeux de données disponibles en ligne. Sa construction nous aura permis de faire face à certaines difficultés techniques, comme la mise en commun de données temporelles disponibles à différentes échelles.

Pour une meilleur reproductibilité de notre travail, nous avons choisi d'une part de développer un package et d'autre part d'écrire l'entièreté de notre projet dans un fichier sweave.

Enfin, ce travail nous aura permis de "mettre les mains dans le cambouis" et d'évoluer dans notre compréhension des problématiques posées par l'étude d'un jeu de données concret. A ce titre, nous remercions Yannig Goude pour son aide précieuse.

Références

[1] Chicago : Midwest Information Office : U.S. Bureau of Labor Statistics.

[2] Joel M. Caplan, Leslie W. Kennedy, and Eric L. Piza. Joint Utility of Event-Dependent and Environmental Crime Analysis Techniques for Violent Crime Forecasting :. *Crime & Delinquency*, November 2012.

[3] Gerhard J. Falk. The Influence of the Seasons on the Crime Rate. *The Journal of Criminal Law, Criminology, and Police Science*, 43(2) :199, July 1952.

[4] Marcus Felson and Erika Poulsen. Simple indicators of crime by time of day. *International Journal of Forecasting*, 19(4) :595–601, 2003.

[5] Wilpen Gorr, Andreas Olligschlaeger, and Yvonne Thompson. Short-term forecasting of crime. *International Journal of Forecasting*, 19(4) :579–594, 2003.

[6] Joel Gunter. Chicago goes high-tech in search of answers to gun crime surge. June 2017.

[7] Simha F. Landau and Daniel Fridman. The Seasonality of Violent Crime : The Case of Robbery and Homicide in Israel. *Journal of Research in Crime and Delinquency*, 30(2) :163–191, May 1993.

[8] John Pepper. *Forecasting Crime : A City Level Analysis*. University of Virginia, Department of Economics, 2007.

[9] Alex Reinhart and Joel Greenhouse. Self-exciting point processes with spatial covariates : modelling the dynamics of crime. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 67(5) :1305–1329, 2018.

[10] National Weather Service Corporate Image Web Team. National Weather Service Climate.

[11] Nick Thieme. Statistics in court. *Significance*, 15(5) :14–17, 2018.

[12] Hongjian Wang, Daniel Kifer, Corina Graif, and Zhenhui Li. Crime Rate Inference with Big Data. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 635–644, New York, NY, USA, 2016. ACM.